A Note on

M-Estimation for Time Series

With an Application to a Simplified FAIR-Model

Ulrich Roschitsch*

November 2022 (last revised February 28, 2024)

`Link to most recent version`

I. Introduction

Many models and estimators in today's time series econometrics rely on nonlinear (or even non-quadratic) optimization problems. In some noteworthy and relevant exceptions, like VAR-models, asymptotically valid inference can be derived from closed-form estimators by applying generic laws of large numbers (LLNs) and central limit theorems (CLTs) for dependent data (E.g. readily available in textbooks like Davidson, 1994; Brockwell and Davis, 2009; Klenke, 2013). By far not all estimators admit such generic treatment, though. The "Functional Approximation of Impulse Responses (FAIR)" estimator of VMA-models proposed by Barnichon and Matthes (2018) is one example, relevant by its recent popularity in the empirical macroeconomics literature (cf., e.g., Barnichon et al., 2021). In short, FAIR proposes to approximate the vector moving average (VMA) representation of a Wold-decomposable process by writing each element of the matrix-lag-polynomial as a sum over finitely many basis functions. Then, the parameters governing shape and weighting of the basis functions in each element of the matrix-lag-polynomial are estimated in a Bayesian fashion in Barnichon and Matthes's original approach. Crucially, while the authors mention that FAIR "can be estimated using maximum likelihood" (Barnichon and Matthes, 2018, p. 9) the asymptotic properties of a such-constructed estimator are not discussed in their paper.

This paper aims to summarize the main results available on the asymptotic properties of M-Estimators for serially dependent data. M-Estimation[1] covers a very broad class of estimators, namely those that can be written as the solution to an in-sample optimization-

*University of Mannheim, Email: ulrich.roschitsch@uni-mannheim.de; Term Paper for: E823 "Advanced Time Series Analysis" by Prof. Carsten Trenkler.

[1] The label is resolved differently, depending on the author(s). White (1996) claims it stands for "Maximum-Likelihood-like" estimation – a reference to the connection between M-Estimation and the (Quasi-) Maxmimum-Likelihood class that he introduced (White, 1982).

problem with an objective that is an average of some terms. Following his introduction of *Quasi*-Maximum-Likelihood (QML) for i.i.d.-data (White, 1982), Halbert White and co-authors dedicated several papers to exploring the asymptotic properties of related estimators with dependent data (Domowitz and White, 1982; White and Domowitz, 1984; Wooldridge, 1986). The results of these efforts, White collected in his 1996 textbook "Estimation, Inference and Specification Analysis", while Wooldridge (1994) provides an overview in the Handbook of Econometrics. This is the main body of work on which this note rests. After examining the general results of this literature, I briefly outline how they may be applied to establish consistency and asymptotic normality of a suitably defined QML-estimator for the class of FAIR-models. It is important to remark here that the task of establishing these properties for the FAIR-estimator on a general VMA-process is somewhat involved – we have (a) an approximation (FAIR vs. VMA) and (b) a latent-variable problem to take care of. I circumvent both issues essentially by assuming them away. This is meant as a first step towards a more general analysis.

In the rest of this paper, I examine the population-level workings of QML (Section II), the general results for consistency (Section III) and asymptotic normality (Section IV) of M-Estimators, as well as a specialized result (Section V): the conditions for the general results can be tremendously sharpened upon focussing on the framework of QML-Estimation, with a Gaussian density, of a correctly specified model (up to second moments). Finally, in Section VI, I apply these sharper results to the simplified FAIR model class.

Notation

As notation, I use the following convention (more specific notation is elaborated on the fly). Scalars: x. Vectors: $\boldsymbol{x}$. Matrices: $\underline{\boldsymbol{x}}$. Element i, j of matrix $\underline{\boldsymbol{x}}$: $[\underline{\boldsymbol{x}}]_{i,j} \in \mathbb{R}$. The converse, a matrix composed of the elements of the doubly array $\mathbb{N}^2 \supset I \times J \to \mathbb{R} : (i, j) \mapsto a_{i,j}$, is denoted $[a_{i,j}]_{i \in I, i \in J}$ (with the understanding that the dimension is $|I| \times |J|$). "vec $\underline{\boldsymbol{X}}$" denotes vectorization of the matrix $\underline{\boldsymbol{X}}$ (the vector obtained from stacking all columns on top of one another, going from left to right); $\otimes$ denotes either the Kronecker product of two matrices or the product of two measures, depending on the context. "$\xrightarrow{\text{p}}$" denotes convergence in probability for a sequence of random variables, "$\xrightarrow{\text{d}}$" denotes convergence in distribution.

II. Quasi-Maximum-Likelihood

Let us first turn to a general description of what Quasi-Maximum-Likelihood estimation (QMLE) does on a population level. This will be helpful in two ways: (a) QMLE is widely applied and has a natural population-level interpretation; and (b) many M-estimators can be reformulated as a QMLE-estimator[2]. White (1996), Ch. 2, delivers a detailed exposition of this subject, which I reproduce up to minor adjustments.

The fundamental inference problem as it pertains to QMLE is as follows: we observe a sample of data, $\{\boldsymbol{x}_t\}_{t=1}^T =: \boldsymbol{x}^T$, $\boldsymbol{x}_t \in \mathbb{R}^k$, and want to learn about features of its distribution, $P_0^T := \mathcal{L}(\boldsymbol{x}^T)$.[3] We make the following

Assumption 1 (Data-generating Process). *The observations $\boldsymbol{x}^T$ are realizations of the stochastic process $(\boldsymbol{x}_t)_{t\in\mathbb{N}}$ on the complete probability space $(\prod_{t\geq 1}\mathbb{R}^k, \otimes_{t\geq 1}\mathcal{B}(\mathbb{R}^k), \mathbb{P}_0)$. Furthermore, for each $T \in \mathbb{N}$, there exists a σ-finite measure ν^T on $(\mathbb{R}^{Tk}, \mathcal{B}(\mathbb{R}^{Tk}))$, known to the analyst and s.th. $P_0^T \ll \nu^T$ with density g^T. (This reflects knowledge of the support of $\boldsymbol{x}^T$.)*

The standard approach is now to specify a set of approximand densities meant to capture g^T, the unknown part of P_0^T. To judge the fit of this approximation, and select a specific ‘best’ approximand, a distance criterion is needed. This is where the intepretability of QMLE is rooted: it judges the fit of a candidate density f by the *Kullback-Leibler*-divergence. This criterion is not formally a metric but possesses many desirable properties, one of which is its information-theoretic interpretation.[4] The general result here is

Definition & Proposition 1 (Kullback-Leibler-Divergence and Information Inequality). *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and let*

(i) $g : (\Omega, \mathcal{A}) \to (\mathbb{R}_{\geq 0}, \mathcal{B}(\mathbb{R}_{\geq 0}))$ *s.th. (a)* $\int_\Omega g \, \mathrm{d}\mu < +\infty$ *and (b)* $\int_S g \log g \, \mathrm{d}\mu < +\infty$ *for* $S := \{\omega \in \Omega : g(\omega) > 0\}$

(ii) $f : (\Omega, \mathcal{A}) \to (\mathbb{R}_{\geq 0}, \mathcal{B}(\mathbb{R}_{\geq 0}))$ *s.th.* $\int_S g \log f \, \mathrm{d}\mu < +\infty$ *(note the absence of* $|\cdot|$*: the value* $\log(0) := -\infty$ *is permitted!)*

Then the Kullback-Leibler-Divergence of f with respect to g is defined as

$$KL(g : f) := \int_S \log[g/f] \cdot g \, \mathrm{d}\mu,$$

and we have the “Information Inequality”:

$$\int_S (g - f) \, \mathrm{d}\mu \geq 0 \implies KL(g : f) \geq 0 \wedge (KL(g : f) = 0 \iff g = f, \ \mu\text{-a.e. on } S).$$

Note that the LHS is always true for probability densities g, f.

[2] By thinking of the summands of the objective as log-density terms; more explanations below.

[3] $\mathcal{L}(\cdot)$ is the operator that associates with each random variable, $X : (\Omega, \mathbb{A}, \mathbb{P}_0) \to (\Omega', \mathbb{A}')$, its “law”, i.e. the unique measure $\mathbb{P} \circ X^{-1}$.

[4] Put loosely, KL measures the information lost when using $f \, \mathrm{d}\nu^T$ as a distribution for $\boldsymbol{x}^T$, rather than $g^T \, \mathrm{d}\nu^T$; it does so by measuring the excess amount of bits needed, on average, to encode a value generated by $g^T \, \mathrm{d}\nu^T$ when using an encryption that is optimized for $f \, \mathrm{d}\nu^T$.

Proof. Since $\int g \,\mathrm{d}\mu < +\infty$, take w.l.o.g. $\int g \,\mathrm{d}\mu \le 1$ and observe that

$$1 \ge \int_S g \,\mathrm{d}\mu \ge \int_S f \,\mathrm{d}\mu = \int_S f/g \cdot g \,\mathrm{d}\mu \overset{\text{Jensen}}{\ge} \exp\left(\int_S \log(f/g) \cdot g \,\mathrm{d}\mu\right)$$

$$\implies 0 \ge \int_S \log(f/g) \cdot g \,\mathrm{d}\mu = -KL(g : f).$$

For $KL(g : f) = 0 \iff g = f$, μ-a.e. on S, sufficiency is obvious. For necessity, assume w.l.o.g. $\mu(S) > 0$ and suppose for contradiction that $\mu(\{g \neq f\} \cap S) > 0$. By $KL = 0$ it must be that $S \equiv F_1 \uplus F_2 \uplus F_3$, with $F_1 := \{g > f\}$, $F_2 := \{g < f\}$, $F_3 := \{g = f\}$, with the first two nonempty.
Now:

$$0 \le \int_S (g - f) \,\mathrm{d}\mu = \int_{F_1} (g - f) \,\mathrm{d}\mu - \int_{F_2} (f - g) \,\mathrm{d}\mu, \quad \text{and}$$

$$\int_S \log(g/f) g \,\mathrm{d}\mu = \int_{F_1} \log(g/f) g \,\mathrm{d}\mu - \int_{F_2} \log(f/g) g \,\mathrm{d}\mu.$$

Furthermore, by the Mean-Value-Theorem, $\forall \omega \in F_2$, $\exists \alpha \in (1, f(\omega)/g(\omega)) : \log(f/g) = (f/g - 1) - (f/g - 1)^2/(2\alpha^2) < (f/g - 1)$. Therefore, $-\int_{F_2} \log(f/g) g \,\mathrm{d}\mu > \int_{F_2} (g - f) \,\mathrm{d}\mu$.

For $\omega \in F_1$ we similarly get $\exists \alpha \in (f(\omega)/g(\omega), 1) : \log(f/g) = (f/g-1)-(f/g-1)^2/(2\alpha^2) < (f/g-1)$ and therefore $\int_{F_1} \log(g/f) g \,\mathrm{d}\mu > \int_{F_1} (g - f) \,\mathrm{d}\mu$. Thus, ultimately:

$$\int_S \log(g/f) g \,\mathrm{d}\mu > \int_{F_1} (g - f) \,\mathrm{d}\mu + \int_{F_2} (g - f) \,\mathrm{d}\mu = \int_S (g - f) \,\mathrm{d}\mu \ge 0 \wedge KL(g : f) = 0 \implies \perp .$$

■

In this KLD, we will plug g^T, f and take $\mu = \nu^T$. But first, we state the density-ratio-decomposition of the population density g^T – this serves as a population counterpart for the familiar forecast-error-factorization in the sample-(quasi-)likelihood function.

Proposition 1 (Density Factorization)**.** *Given the process in A1, some $T \in \mathbb{N}$, and the true law $P_0^T \ll \nu^T$, it is possible to choose a version of the density, g^T, s.th.*

$$\boldsymbol{x}^T \in S^T := \{\boldsymbol{x}^T : g^T(\boldsymbol{x}^T) > 0\} \implies \boldsymbol{x}^{T-1} \in S^{T-1}$$

(such g^T are called "standard") so that the following the following holds P_0^T-a.s.:

$$\log g^T(\boldsymbol{x}^T) = \sum_{t=1}^{T} \log g_t(\boldsymbol{x}^t), \quad g_t(\boldsymbol{x}^t) := \frac{g^t(\boldsymbol{x}^t)}{g^{t-1}(\boldsymbol{x}^{t-1})}.$$

(If $\nu^T \equiv \bigotimes_{t=1}^T \nu_t$, g_t is a conditional density.)

This decomposition has the significant advantage that we can define approximands directly for the density ratio g_t (Note that g_t may still depend on T – for notational convenience, this is left implicit). This setup specializes very nicely to cases where $g_t \equiv g$. Now we can finally define our approximands:

Definition 2 (Parametric Stochastic Specification)**.** *Grant Assumption 1 and consider the measure* $g^T \, \mathrm{d}\nu^T$, $T \in \mathbb{N}$. *A parametric stochastic specification for* $(g_t) := \{g^T, T \in \mathbb{N}\}$ *is a triangular array of functions* $(f_t) = \{(f_t)_{t=1}^T, T \in \mathbb{N}\}$ *where* $\forall T$, $f_t : \mathbb{R}^{tk} \times \Theta \to \mathbb{R}_{\geq 0}$ *with* $\Theta \subseteq \mathbb{R}^p$ *and s.th.* $\forall \boldsymbol{\theta} \in \Theta, f_t(\cdot, \boldsymbol{\theta})$ *is measurable.*

The basic idea here is that $P_{\boldsymbol{\theta}}^T : \mathcal{B}(\mathbb{R}^{Tk}) \ni B \mapsto \int_{B \cap S^T} \prod_{t=1}^T f_t(\boldsymbol{x}^t, \boldsymbol{\theta}) \, \mathrm{d}\nu^T$ is a probability measure that approximates P_0^T well in the sense that KL is small. However, while $P_{\boldsymbol{\theta}}^T$ is indeed a well-defined measure, it need not be a probability measure. Even if it is, it need then not be that f_t is a conditional density, even if g_t is. These facts are especially relevant for cases where the approximands f_t specify conditional distribution features of a sub-vector of $\boldsymbol{x}_t$, given another sub-vector of it – in this case, f_t evidently does not pin down a full distribution of $\boldsymbol{x}_t$.

Beyond the previous requirements, the following regularity conditions are helpful.

Assumption 2 (Regularity of f_t)**.** (f_t) *is s.th.* $\forall t \leq T \in \mathbb{N}$,

(i) $\forall \boldsymbol{\theta} \in \Theta, f_t(\cdot, \boldsymbol{\theta})$ *is measurable and* $\int_{S^T} \prod_{t=1}^T f_t(\boldsymbol{x}^t, \boldsymbol{\theta}) \, \mathrm{d}\nu^T \leq 1$,[5]

(ii) $f_t(\boldsymbol{x}^T, \cdot)$ *is* P_0^T*-a.s. continuous.*

Now for the population-level program of QMLE: fix some $T \in \mathbb{N}$ and (f_t). Then the **"quasi-likelihood-function"**[6] $\boldsymbol{\theta} \mapsto \prod_{t=1}^T f_t =: f^T$ can be viewed as an approximation to $g^T = \prod_{t=1}^T g_t$ with adequacy measured by

$$KL(g^T : f^T(\cdot, \boldsymbol{\theta})) = \int_{S^T} \log[g^T / f^T] g^T \, \mathrm{d}\nu^T, \text{ which is minimized by maximizing}$$

$$T \cdot \bar{L}_T(\boldsymbol{\theta}) := \int_{S^T} \log[f^T(\boldsymbol{x}^T, \boldsymbol{\theta})] g^T(\boldsymbol{x}^T) \nu^T(\mathrm{d}\boldsymbol{x}^T).$$

That is, given a class of approximands (f_t), we can find the KL-best approximand by solving

$$\max_{\boldsymbol{\theta}} \bar{L}_T(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^T \mathbb{E}\{\log f_t(\boldsymbol{x}^t, \boldsymbol{\theta})\}. \tag{II.1}$$

If $\exists \boldsymbol{\theta}_0 \in \Theta : f^T(\cdot, \boldsymbol{\theta}_0) = g^T$, P_0^T-a.s., then by the information inequality and A2 the above program has the **unique solution** $\boldsymbol{\theta}_0$.

[5] This qualification is not made by White (1996), but is important so that the information inequality may be applied later.

[6] Actually, this label is traditionally used for a sample context; White (1996) uses it to label the random function given here.

The link to the sample-definition of QMLE is now straightforward. Since we lack knowledge of g^T we cannot evaluate $\bar{L}_T(\cdot)$. The QMLE-approach proceeds to use sample-information and a LLN[7] to elicit the $\bar{L}_T$-maximizer from the **"sample quasi-likelihood"**:

$$L_T(\boldsymbol{\theta}) := \frac{1}{T}\sum_{t=1}^{T} \log f_t(\boldsymbol{x}^t, \boldsymbol{\theta}). \tag{II.2}$$

The maxmizer (if existent and measurable) is called the **QML-estimator**, $\hat{\boldsymbol{\theta}}$. That the QMLE exists and is measurable is assured by our Assumptions 1 and 2, the requirement Θ be compact and the following

Lemma 1 (Existence and Measurability of a Maximizer). *Let $(\Omega, \mathcal{A})$ be a measurable space, $\Omega \subseteq \mathbb{R}^p$ compact and $Q : \Omega \times \Theta \to \overline{\mathbb{R}}$ s.th. $\forall \boldsymbol{\theta} \in \Theta$, $Q(\cdot, \boldsymbol{\theta})$ is measurable and $\exists A \in \mathcal{A} : \forall \omega \in A$, $Q(\omega, \cdot)$ is continuous. Then, $\exists\, \hat{\boldsymbol{\theta}} : \Omega \to \Theta$, $\mathcal{A}/\mathcal{B}(\Theta)$-measurable and s.th.*

$$\forall \omega \in A,\ Q(\omega, \hat{\boldsymbol{\theta}}(\omega)) = \max_{\boldsymbol{\theta}\in\Theta} Q(\omega, \boldsymbol{\theta}).$$

III. Consistency

Armed with an understanding of how QMLE works at the population level, we can turn to the asymptotic properties of general M-Estimators, the first step of which is to study general results for consistency. Wooldridge (1994) gives a good synthesis of the literature, outlining most of the herein presented results. M-Estimators are generally defined as

$$\hat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\Theta} \underbrace{\frac{1}{T}\sum_{t=1}^{T} q_t(\boldsymbol{w}_t, \boldsymbol{\theta})}_{=:\widehat{Q}_T(\boldsymbol{\theta})}, \tag{III.1}$$

where $\Theta \subseteq \mathbb{R}^p$ is the parameter space, $q_t : \mathcal{W}_t \times \Theta \to \mathbb{R}$ is a function with $\boldsymbol{w}_t \in \mathcal{W}_t \subseteq \mathbb{R}^{k_t}$ some random vector with the dimension growing with t usually, $k_t \to \infty$, as $t \to \infty$.[8]

To develop a consistency notion, we require a target-parameter (or generally a sequence) against which $\hat{\boldsymbol{\theta}}$ is supposed to be consistent. As for the case of QMLE, this consistency

[7] Recall that for a sequence of random variables $(Y_t)_{t\in\mathbb{N}}$ a LLN holds if $\frac{1}{T}\sum_{t=1}^{T}(Y_t - \mathbb{E}Y_t) \xrightarrow{\text{p}} 0$.

[8] All necessary discussions on existence, measurability and uniqueness of such $\hat{\boldsymbol{\theta}}$ will be handled inside the theorems.

target is taken to be the minimizer of a suitable population counterpart of the objective (III.1):

$$(\boldsymbol{\theta}_T^*)_{T\in\mathbb{N}}, \quad \text{where } \forall T \in \mathbb{N},\, \boldsymbol{\theta}_T^* = \arg\min_{\boldsymbol{\theta}\in\Theta} \underbrace{\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\{q_t(\boldsymbol{w}_t,\boldsymbol{\theta})\}}_{=:\overline{Q}_T(\boldsymbol{\theta})}. \tag{III.2}$$

The first theorem on consistency in the sense that $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_T^* \xrightarrow{\text{p}} 0$ is from White (1996), Ch. 3:

Theorem 2 (Consistency with heterogeneous population-objectives)**.** *Consider $\hat{\boldsymbol{\theta}}$ as above and suppose the following holds:*

(A0) $\Theta \subset \mathbb{R}^p$ *compact,*

(A1) $\boldsymbol{w}_t = \boldsymbol{x}^t$*, where* $(\boldsymbol{x}_t)_{t\in\mathbb{N}}$ *satisfies Assumption 1 and* $f_t \equiv \exp(-q_t)$ *satisfies Assumption 2 (this is the link between QMLE and M-Estimation),*

(A2) $\forall \boldsymbol{\theta} \in \Theta$, $\forall t \leq T \in \mathbb{N}$*: (a)* $\mathbb{E}q_t(\boldsymbol{x}^t,\boldsymbol{\theta})$ *exists in* $\mathbb{R}$*, (b)* $\boldsymbol{\theta} \mapsto \mathbb{E}q_t(\boldsymbol{x}^t,\boldsymbol{\theta})$ *is continuous,*[9] *(c)* $\boldsymbol{\theta} \mapsto \overline{Q}_T(\boldsymbol{\theta})$ *has identifiably unique minimizers in the limit: (given* Θ *compact)* $\forall \varepsilon > 0$, $\limsup_{T\to\infty} \min_{\boldsymbol{\theta}\in(B_\varepsilon(\boldsymbol{\theta}_T^*))^{\mathrm{c}}\cap\Theta}(\overline{Q}_T(\boldsymbol{\theta}) - \overline{Q}_T(\boldsymbol{\theta}_T^*)) > 0$*,*

(A3) $q_t(\cdot,\cdot)$ *obeys the weak uniform law of large numbers (WULLN), i.e.*

$$\max_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{T} q_t(\boldsymbol{w}_t,\boldsymbol{\theta}) - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{q_t(\boldsymbol{w}_t,\boldsymbol{\theta})\}\right| \xrightarrow{\text{p}} 0.$$

Then, there exists a measurable minimizer $\hat{\boldsymbol{\theta}} : \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_T^* \xrightarrow{\text{p}} 0$*.*

As reassuring as the availability of such a general theorem is, in many practical applications a simpler result (see Wooldridge, 1994) can be used, the (quite short) proof of which gives a good deal of insight into the workings of these consistency results. The theorem presupposes that there exists not only a sequence of targets $(\boldsymbol{\theta}_T^*)$, but a well-defined limit-function with associated minimizer:

$$\exists \boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} \underbrace{\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{q_t(\boldsymbol{w}_t,\boldsymbol{\theta})\}}_{=:\overline{Q}(\boldsymbol{\theta})} \tag{III.3}$$

which is, e.g., satisfied if $(q_t(\boldsymbol{w}_t,\boldsymbol{\theta}))$ is stationary with a unique minimizer (more below).

Theorem 3 (Consistency with limit population-objective)**.** *Consider $\hat{\boldsymbol{\theta}}$ as above and suppose Theorem 2, (A0)&(A1) hold. Suppose further:*

[9] This is guaranteed by (A1).

(A2)' *(a)* $\overline{Q} : \Theta \to \mathbb{R}$ *defined above exists, (b)* $\overline{Q}(\cdot)$ *has the identifiably unique minimizer* $\boldsymbol{\theta}^*$*, i.e.* $\forall \varepsilon > 0$, $\min_{\boldsymbol{\theta}\in(B_\varepsilon(\boldsymbol{\theta}_T^*))^{\complement}\cap\Theta} \overline{Q}(\boldsymbol{\theta}) > \overline{Q}(\boldsymbol{\theta}^*)$,

(A3)' $q_t(\cdot,\cdot)$ *obeys the WULLN, i.e.* $\max_{\boldsymbol{\theta}\in\Theta} \left|\widehat{Q}_T(\boldsymbol{\theta}) - \overline{Q}(\boldsymbol{\theta})\right| \xrightarrow{\text{p}} 0.$

Then, there exists a measurable minimizer $\hat{\boldsymbol{\theta}} \xrightarrow{\text{p}} \boldsymbol{\theta}^*$.

Proof. Existence and measurability of $\hat{\boldsymbol{\theta}}$ follow from Lemma 1; uniquness is not needed for consistency, so the notation "= arg min" is left as abuse of notation. It is required to show $\forall \varepsilon > 0$, $\lim_{T\to\infty} \mathbb{P}\left(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| > \varepsilon\right) = 0$. Fix $\varepsilon > 0$ and note that by (A2)'-(b), a $\delta > 0$ exists, s.th. $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| > \varepsilon \implies |\overline{Q}(\hat{\boldsymbol{\theta}}) - \overline{Q}(\boldsymbol{\theta}^*)| > \delta$, whence it follows $\mathbb{P}\left(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| > \varepsilon\right) \leq \mathbb{P}(|\overline{Q}(\hat{\boldsymbol{\theta}}) - \overline{Q}(\boldsymbol{\theta}^*)| > \delta)$. So it is sufficient to show $\overline{Q}(\hat{\boldsymbol{\theta}}) \xrightarrow{\text{p}} \overline{Q}(\boldsymbol{\theta}^*)$. To this end, one can make the estimations

$$\begin{aligned}
0 \leq \overline{Q}(\hat{\boldsymbol{\theta}}) - \overline{Q}(\boldsymbol{\theta}^*) &= \overline{Q}(\hat{\boldsymbol{\theta}}) - \widehat{Q}_T(\hat{\boldsymbol{\theta}}) + \widehat{Q}_T(\hat{\boldsymbol{\theta}}) - \overline{Q}(\boldsymbol{\theta}^*) \\
&\leq \overline{Q}(\hat{\boldsymbol{\theta}}) - \widehat{Q}_T(\hat{\boldsymbol{\theta}}) + \widehat{Q}_T(\boldsymbol{\theta}^*) - \overline{Q}(\boldsymbol{\theta}^*) \quad \text{by } \boldsymbol{\theta}^* = \arg\min \\
&\leq 2 \max_{\boldsymbol{\theta}\in\Theta} \left|\widehat{Q}_T(\boldsymbol{\theta}) - \overline{Q}(\boldsymbol{\theta})\right| \xrightarrow{\text{p}} 0.
\end{aligned}$$

■

The key enabling assumption in both theorems is the WULLN. Wooldridge (1994) presents two results on how to establish a WULLN, one for a time-heterogeneous setup as above, and a more specialized 'homogeneous' result (which will be sufficient for our purposes later):

Lemma 2 (Homogeneous WULLN). *Let*

(i) $(\boldsymbol{w}_t)_{t\in\mathbb{N}}$ *be a strictly stationary and ergodic*[10] *process on* $(\mathcal{W}^\infty, \mathcal{B}(\mathcal{W})^{\otimes\infty})$, $\mathcal{W} \subseteq \mathbb{R}^k$

(ii) $\Theta \subseteq \mathbb{R}^p$ *be compact*

(iii) $q : \mathcal{W} \times \Theta \to \mathbb{R}$ *be s.th.* $q(\cdot, \boldsymbol{\theta})$ *measurable* $\forall \boldsymbol{\theta} \in \Theta$*, and* $q(\boldsymbol{w}_t, \cdot)$ *a.s. continuous*

(iv) $\exists b : \mathcal{W} \to \mathbb{R}$ *measurable and s.th.* $\mathbb{E}b < +\infty$ *and* $|q(\boldsymbol{w}_t, \boldsymbol{\theta})| \leq b(\boldsymbol{w}_t)$ *a.s.* $\forall t, \forall \boldsymbol{\theta}$

Then, $\max_{\boldsymbol{\theta}\in\Theta} \left|\frac{1}{T}\sum_{t=1}^T \left(q(\boldsymbol{w}_t, \boldsymbol{\theta}) - \mathbb{E}\{q(\boldsymbol{w}_t, \boldsymbol{\theta})\}\right)\right| \xrightarrow{\text{p}} 0.$

Proof. By (i) and (iii), the process $(q(\boldsymbol{w}_t, \boldsymbol{\theta}))_{t\in\mathbb{N}}$ is strictly stationary and ergodic for any $\boldsymbol{\theta}$, hence we have a pointwise LLN by Birkhoff's ergodic theorem (e.g. Klenke, 2013, Theorem 20.14). Furthermore, conditions (ii), (iii) and (iv) are sufficient for $\boldsymbol{\theta} \mapsto \frac{1}{T}\sum_{t=1}^T q(\boldsymbol{w}_t, \boldsymbol{\theta})$ to be stochastically equicontinous (SEQ).[11] SEQ follows here in three steps:

[10] Weakly stationary would also suffice, given absolute summability of the autocovariances of the ensuing $q(\boldsymbol{w}_t, \boldsymbol{\theta})$.

[11] A sequence of random functions $(\widehat{Q}_T)_{T\in\mathbb{N}}$, $\widehat{Q}_T : \Omega \times \Theta \to \mathbb{R}$ is SEQ if: $\forall \varepsilon > 0 \exists \delta > 0 : \limsup_{T\to\infty} \mathbb{P}(w(\widehat{Q}_T, \delta) \geq \varepsilon) < \varepsilon$, where $w(\widehat{Q}_T, \delta) := \sup_{\boldsymbol{\theta}\in\Theta} \sup_{\boldsymbol{\vartheta}\in B_\delta(\boldsymbol{\theta})\cap\Theta} |\widehat{Q}_T(\boldsymbol{\theta}) - \widehat{Q}_T(\boldsymbol{\vartheta})|$ is the "modulus of continuity".

1) $w(\widehat{Q}_T, \delta) \xrightarrow{\text{a.s.}} 0$ as $\delta \to 0$: pick some $\bar{\omega} \in \mathcal{C} := \{\omega \in \mathcal{W}^\infty : \widehat{Q}_T(\omega, \cdot) \text{ is continuous}\}$ and note that $\mathbb{P}(\mathcal{C}) = 1$ by (iii); now since Θ is compact, $\boldsymbol{\theta} \mapsto \widehat{Q}_T(\bar{\omega}, \boldsymbol{\theta})$ is uniformly continuous; thus, choosing some $\eta > 0$, this ensures $\exists \delta > 0 : \forall \boldsymbol{\theta}, \boldsymbol{\vartheta} \in \Theta,\ \|\boldsymbol{\theta} - \boldsymbol{\vartheta}\| < \delta \implies |\widehat{Q}_T(\boldsymbol{\theta}) - \widehat{Q}_T(\boldsymbol{\vartheta})| < \eta$.

2) By (iv), $w(\widehat{Q}_T, \delta) \leq 2\frac{1}{T}\sum_{t=1}^T b(\boldsymbol{w}_t)$ and by dominated convergence, $\mathbb{E}\{w(\widehat{Q}_T, \delta)\} \to 0$ as $\delta \to 0$.

3) Apply Markov's inequality to conclude $\exists \delta > 0 : \mathbb{P}(w(\widehat{Q}_T, \delta) \geq \varepsilon) < \varepsilon$ as $T \to \infty$ for any $\varepsilon > 0$.

These are all the conditions needed to apply Theorem 21.9 by Davidson (1994). ■

IV. Normality

For providing distributional approximations for the estimator $\hat{\boldsymbol{\theta}}$, results on asymptotic normality are the usual option. While White (1996) provides results for the fully time-heterogeneous case ($\boldsymbol{\theta}_T^* \nrightarrow \boldsymbol{\theta}^*$) in Ch. 6.1, I concentrate on Wooldridge's (1994) exposition which relies on the existence of the time-invariant population program (III.3).

Wooldridge's main result exploits, as in the cross-section-case, the mean-value decomposition of the FOC to the (differentiable) program (III.1). Thus, it relies on a CLT for the score vector (the derivative of $q_t(\boldsymbol{w}_t, \boldsymbol{\theta})$ for $\boldsymbol{\theta}$); sufficient conditions for this CLT are given in the next section.

Theorem 4 (Asymptotic Normality of M-Estimators). *Grant the assumptions (A0)–(A3)' of Theorem 3 and suppose additionally that*

(A4) (a) $\boldsymbol{\theta}^* \in \operatorname{int}\Theta$, (b) $(\boldsymbol{\theta} \mapsto q_t(\cdot, \boldsymbol{\theta})) \in C^2(\operatorname{int}\Theta)$ a.s. $\forall t \in \mathbb{N}$; define the score and Hessian (summand)

$$\boldsymbol{s}_t(\boldsymbol{w}_t, \boldsymbol{\theta}) := \frac{\partial q_t(\boldsymbol{w}_t, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \quad \underline{\boldsymbol{h}}_t(\boldsymbol{w}_t, \boldsymbol{\theta}) := \frac{\partial^2 q_t(\boldsymbol{w}_t, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}.$$

(A5) (a) $(\underline{\boldsymbol{h}}_t(\boldsymbol{w}_t, \boldsymbol{\theta}))_{t \in \mathbb{N}}$ satisfies the WULLN on Θ with $\underline{\boldsymbol{A}} := \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \mathbb{E}\{\underline{\boldsymbol{h}}_t(\boldsymbol{w}_t, \boldsymbol{\theta}^*)\}$, (b) $(\boldsymbol{s}_t(\boldsymbol{w}_t, \boldsymbol{\theta}^*))_{t \in \mathbb{N}}$ satisfies a CLT, i.e.:

$$\sqrt{T}\frac{1}{T}\sum_{t=1}^T \boldsymbol{s}_t(\boldsymbol{w}_t, \boldsymbol{\theta}^*) \xrightarrow{\text{d}} \mathcal{N}(\boldsymbol{0}, \underline{\boldsymbol{\Xi}}),$$

$$\underline{\boldsymbol{\Xi}} := \lim_{T\to\infty} \operatorname{Var}\left(\sqrt{T}\frac{1}{T}\sum_{t=1}^T \boldsymbol{s}_t(\boldsymbol{w}_t, \boldsymbol{\theta}^*)\right) >_{\text{p.d.}} \underline{\boldsymbol{0}}$$

Then, $\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{\text{d}} \mathcal{N}(\boldsymbol{0}, \underline{\boldsymbol{A}}^{-1}\underline{\boldsymbol{\Xi}}\underline{\boldsymbol{A}}^{-1})$.

Proof. Because of (A4) and $\hat{\boldsymbol{\theta}} \xrightarrow{\text{p}} \boldsymbol{\theta}^*$, $\hat{\boldsymbol{\theta}}$ solves $\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{s}_t(\boldsymbol{w}_t, \hat{\boldsymbol{\theta}}) = 0$, with probability approaching 1.[12] Using the Mean-Value-Theorem, we may write for some $\tilde{\boldsymbol{\theta}} \in (\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$,

$$\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{s}_t(\boldsymbol{w}_t, \boldsymbol{\theta}^*) + \left(\frac{1}{T}\sum_{t=1}^{T} \underline{\boldsymbol{h}}_t(\boldsymbol{w}_t, \tilde{\theta})\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \mathbf{0}$$

$$\iff \sqrt{T}\underbrace{\left(\frac{1}{T}\sum_{t=1}^{T} \underline{\boldsymbol{h}}_t(\boldsymbol{w}_t, \tilde{\theta})\right)}_{\text{Will be invertible with probability approaching 1}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \sqrt{T}\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{s}_t(\boldsymbol{w}_t, \boldsymbol{\theta}^*)$$

$$\iff \sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \left(\frac{1}{T}\sum_{t=1}^{T} \underline{\boldsymbol{h}}_t(\boldsymbol{w}_t, \tilde{\theta})\right)^{-1} \sqrt{T}\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{s}_t(\boldsymbol{w}_t, \boldsymbol{\theta}^*).$$

Finally, using Slutsky's lemma and the WULLN in combination with the stochastic sandwich theorem delivers the result. ■

If such an asymptotic normality result holds in a given application, consistent variance estimation of $\text{Avar}(\hat{\boldsymbol{\theta}}) = \underline{\boldsymbol{A}}^{-1}\underline{\boldsymbol{\Xi}}\underline{\boldsymbol{A}}^{-1}$ may be achieved by suitable plug-in estimators. By the conditions from the theorem, $\underline{\boldsymbol{A}} = \text{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \underline{\boldsymbol{h}}_t(\boldsymbol{w}_t, \hat{\boldsymbol{\theta}})$, while consistent estimation of $\underline{\boldsymbol{\Xi}}$ is highly case-dependent. E.g. for a weakly stationary score (with finite 4th moments) the usual HAC-machinery can be applied to estimate $\underline{\boldsymbol{\Xi}} = \underline{\boldsymbol{\Gamma}}_0 + \sum_{h\in\mathbb{N}}(\underline{\boldsymbol{\Gamma}}_h + \underline{\boldsymbol{\Gamma}}_h^\top)$, $\underline{\boldsymbol{\Gamma}}_h := \mathbb{E}\{\boldsymbol{s}_{t+h}(\boldsymbol{w}_{t+h}, \boldsymbol{\theta}^*)\boldsymbol{s}_t(\boldsymbol{w}_t, \boldsymbol{\theta}^*)^\top\}$.

V. $\mathcal{N}$-QMLE with Correctly Specified Moments

As usual in the theoretic time series literature, primitive conditions for strong theorems are obtained only in exchange for being more specific about the environment. The previous results are no exception. However, in a well-cited[13] contribution, Bollerslev and Wooldridge (1992) manage to sharpen the above results quite a bit, by considering the practically relevant use-case of QMLE-estimation, with a Gaussian density, of dynamic models under the provision that the first and second conditional moments be correctly specified.

The environment can be sketched as follows. Let $(\boldsymbol{y}_t)_{t\in\mathbb{N}}$, $\boldsymbol{y}_t \in \mathbb{R}^{\kappa_1}$ and $(\boldsymbol{z}_t)_{t\in\mathbb{N}}$, $\boldsymbol{z}_t \in \mathbb{R}^{\kappa_2}$ be stochastic processes (outcome and regressor, respectively) and define some conditioning vector $\boldsymbol{x}_t \subseteq (\boldsymbol{z}_t^\top, \boldsymbol{y}_{t-1}^\top, \boldsymbol{z}_{t-1}^\top, \ldots, \boldsymbol{y}_1^\top, \boldsymbol{z}_1^\top)^\top$ and $\boldsymbol{w}_t := (\boldsymbol{y}_t^\top, \boldsymbol{x}_t^\top)^\top \in \mathbb{R}^{k_t}$. Suppose $\exists! \boldsymbol{\theta}_0 \in \Theta \subseteq \mathbb{R}^p$ compact s.th.

$$\forall t \in \mathbb{N}, \quad \mathbb{E}\{\boldsymbol{y}_t | \boldsymbol{x}_t\} \equiv \boldsymbol{\mu}_t(\boldsymbol{x}_t, \boldsymbol{\theta}_0), \quad \text{Var}\{\boldsymbol{y}_t | \boldsymbol{x}_t\} \equiv \underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t, \boldsymbol{\theta}_0).$$

[12] The statement "with probability approaching 1" is used for streamlining. It should rigorously be interchanged with the subsequence theorem ($x_n \xrightarrow{\text{p}} x \iff \exists x_{n_k} \xrightarrow{\text{a.s.}} x$). That is, whenever "w.p.a.1" appears, we implicitly take a subsequence which satisfies the claim on an almost-sure event.

[13] 4,329 citations on Google Scholar (Jan 10, 2022).

The relevant estimator is taken to be the Quasi Maximum-Likelihood-estimator with Gaussian density ($\mathcal{N}$-QMLE),

$$\hat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta}\in\Theta}{\arg\max}\, \frac{1}{T}\sum_{t=1}^{T} \log f_t(\boldsymbol{w}_t, \boldsymbol{\theta}), \tag{V.1}$$

$$\log f_t(\boldsymbol{w}_t,\boldsymbol{\theta}) := -\frac{1}{2}\cdot\big[\kappa_1 \log(2\pi) + \log\det \underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t,\boldsymbol{\theta}) \\ +(\boldsymbol{y}_t - \boldsymbol{\mu}_t(\boldsymbol{x}_t,\boldsymbol{\theta}))^\top \underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t,\boldsymbol{\theta})^{-1}(\boldsymbol{y}_t - \boldsymbol{\mu}_t(\boldsymbol{x}_t,\boldsymbol{\theta}))\big].$$

Then Bollerslev and Wooldridge (1992) proceed to show:

Proposition 5. *If $\underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t,\boldsymbol{\theta}) >_{\text{p.d.}} \underline{\boldsymbol{0}}$, $\forall \boldsymbol{\theta}\in\Theta$, then $\boldsymbol{\theta}_0$ is the identifiably unique maximizer of $\boldsymbol{\theta}\mapsto \mathbb{E}\{\log f_t(\boldsymbol{w}_t,\boldsymbol{\theta})\}$, $\forall t\in\mathbb{N}$.*

Proof. Using the intelligent zero $\boldsymbol{y}_t - \boldsymbol{\mu}_t(\boldsymbol{x}_t,\boldsymbol{\theta}) = \boldsymbol{y}_t - \boldsymbol{\mu}_t(\boldsymbol{x}_t,\boldsymbol{\theta}_0) + \boldsymbol{\mu}_t(\boldsymbol{x}_t,\boldsymbol{\theta}_0) - \boldsymbol{\mu}_t(\boldsymbol{x}_t,\boldsymbol{\theta})$ and commutativity inside the tr-operator, we get

$$\begin{aligned} 2\cdot\mathbb{E}\{\log f_t(\boldsymbol{w}_t,\boldsymbol{\theta})\} &= -\log\det \underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t,\boldsymbol{\theta}) - \operatorname{tr}\big[\underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t,\boldsymbol{\theta})^{-1}\underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t,\boldsymbol{\theta}_0)\big] \\ &\quad - (\boldsymbol{\mu}_t(\boldsymbol{x}_t,\boldsymbol{\theta}_0) - \boldsymbol{\mu}_t(\boldsymbol{x}_t,\boldsymbol{\theta}))^\top \underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t,\boldsymbol{\theta})^{-1}(\boldsymbol{\mu}_t(\boldsymbol{x}_t,\boldsymbol{\theta}_0) - \boldsymbol{\mu}_t(\boldsymbol{x}_t,\boldsymbol{\theta})) \\ &=: Q(\boldsymbol{\mu}_t(\boldsymbol{x}_t,\boldsymbol{\theta}), \underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t,\boldsymbol{\theta})), \end{aligned}$$

and we see $Q(\boldsymbol{\mu}_t(\boldsymbol{x}_t,\boldsymbol{\theta}_0), \underline{\boldsymbol{H}}) > Q(\boldsymbol{m}, \underline{\boldsymbol{H}})$, $\forall \underline{\boldsymbol{H}} >_{\text{p.d.}} \underline{\boldsymbol{0}}$, $\forall \boldsymbol{m} \neq \boldsymbol{\mu}_t(\boldsymbol{x}_t,\boldsymbol{\theta}_0)$. We can also show $Q(\boldsymbol{\mu}_t(\boldsymbol{x}_t,\boldsymbol{\theta}_0), \underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t,\boldsymbol{\theta}_0)) > Q(\boldsymbol{\mu}_t(\boldsymbol{x}_t,\boldsymbol{\theta}_0), \underline{\boldsymbol{H}})$, $\forall \underline{\boldsymbol{H}} \neq \underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t,\boldsymbol{\theta}_0) : \underline{\boldsymbol{H}} >_{\text{p.d.}} \underline{\boldsymbol{0}}$. This is achieved by

$$\begin{aligned} &\log\det \boldsymbol{\mu}_t(\boldsymbol{x}_t,\boldsymbol{\theta}_0) + \operatorname{tr}\underline{\boldsymbol{I}} < \log\det\underline{\boldsymbol{H}} + \operatorname{tr}\underline{\boldsymbol{H}}^{-1}\underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t,\boldsymbol{\theta}_0) \\ \iff\ &\log\det \underline{\boldsymbol{H}}^{-1}\underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t,\boldsymbol{\theta}_0) < \operatorname{tr}\big[\underline{\boldsymbol{H}}^{-1}\underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t,\boldsymbol{\theta}_0) - \underline{\boldsymbol{I}}\big], \end{aligned}$$

which holds for all positive definite matrices (cf. Magnus and Neudecker, 1988, Theorem 27 & following corollaries). The identifiability follows from continuity and compactness of Θ. ■

This directly implies that $\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}\in\Theta} \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{\log f_t(\boldsymbol{w}_t,\boldsymbol{\theta})\}$, $\forall T\in\mathbb{N}$, and by the appropriate assumptions on $\mathcal{L}\left((\boldsymbol{y}_t, \boldsymbol{z}_t)_{t\in\mathbb{N}}\right)$, Theorem 2 ensures that $\hat{\boldsymbol{\theta}} \xrightarrow{\text{p}} \boldsymbol{\theta}_0$. This leaves essentially only the WULLN to the specific application.

We can also make substantial progress when it comes to asymptotic normality. To this end, grant the appropriate differentiability assumptions to $\boldsymbol{\mu}_t$, $\underline{\boldsymbol{\Omega}}_t$ and observe that

$$\boldsymbol{s}_t(\boldsymbol{w}_t,\boldsymbol{\theta}) = \underset{(p\times 1)}{\frac{\partial}{\partial\boldsymbol{\theta}}}\ \log f_t(\boldsymbol{w}_t,\boldsymbol{\theta}) \qquad \text{defining } \boldsymbol{u}_t(\boldsymbol{x}_t,\boldsymbol{\theta}) := \boldsymbol{y}_t - \boldsymbol{\mu}_t(\boldsymbol{x}_t,\boldsymbol{\theta}),$$

$$= \frac{\partial\boldsymbol{\mu}_t(\boldsymbol{x}_t,\boldsymbol{\theta})^\top}{\partial\boldsymbol{\theta}}\underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t,\boldsymbol{\theta})^{-1}\boldsymbol{u}_t(\boldsymbol{x}_t,\boldsymbol{\theta}) - \frac{1}{2}\frac{\partial(\operatorname{vec}(\underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t,\boldsymbol{\theta})))^\top}{\partial\boldsymbol{\theta}}\left(\underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t,\boldsymbol{\theta})^{-1}\otimes\underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t,\boldsymbol{\theta})^{-1}\right)$$

$$\cdot \operatorname{vec}\left[\boldsymbol{u}_t(\boldsymbol{x}_t,\boldsymbol{\theta})\boldsymbol{u}_t(\boldsymbol{x}_t,\boldsymbol{\theta})^\top - \underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t,\boldsymbol{\theta})\right]. \tag{V.2}$$

(A full derivation is provided in the Appendix.) From this we can see that $\mathbb{E}\{\boldsymbol{s}_t(\boldsymbol{w}_t,\boldsymbol{\theta}_0)\,|\,\boldsymbol{x}_t\} = \mathbf{0}$. Therefore, provided we have a condition that Bollerslev and Wooldridge (1992) refer to as "dynamic completeness"

$$\sigma\left(\boldsymbol{s}_\tau(\boldsymbol{w}_t,\boldsymbol{\theta}_0),\ \tau \in \{1,\ldots,t-1\}\right) \overset{\boldsymbol{s}_t(\cdot,\boldsymbol{\theta}_0)\text{ m'able}}{\subseteq} \sigma\left((\boldsymbol{y}_\tau^\top,\boldsymbol{x}_\tau^\top)^\top,\ \tau \in \{1,\ldots,t-1\}\right) \subseteq \sigma(\boldsymbol{x}_t)$$

$$(\Longleftarrow \boldsymbol{x}_t = (\boldsymbol{z}_t^\top,\boldsymbol{y}_{t-1}^\top,\boldsymbol{z}_{t-1}^\top,\ldots,\boldsymbol{y}_1^\top,\boldsymbol{z}_1^\top)^\top,\ \text{e.g.})$$

the law of iterated expectations[14] and the above nullity result ensure that $(\boldsymbol{s}_t(\boldsymbol{w}_t,\boldsymbol{\theta}_0))_{t\in\mathbb{N}}$ is a Martingale Difference Sequence (MDS) with respect to any filtration adapted to it. This means we can, once the appropriate side conditions are established, use an MDS-CLT (see, e.g., White, 2014, p. 130f.) to establish (A5)-(b) in Theorem 4.

VI. Application to a Simple FAIR-Model

Finally, we can apply the above results to a specific class of models and estimation procedures: the "Functional Approximation of Impulse Responses (FAIR)" estimator of VMA-models proposed by Barnichon and Matthes (2018). In fact, all derivations in this section apply generally to VMA-models with finitely many parameters, since I do not explicitly exploit the functional form of the mapping from deep parameters, $\boldsymbol{\theta}_0$, to MA-coefficients. That notwithstanding, the results given here are severely limited in scope by two rather restrictive assumptions that make the framework 'simple'. The results should therefore be taken as a first step towards deriving asymptotics for the FAIR-estimator in a (more) general setting.
Overall, FAIR-estimation of a VMA-model entails two main complications:

(1) Approximation problem: If $(\boldsymbol{y}_t) \sim \text{VMA}(\infty)$ as is often implied by linearized DSGE-models, $(\boldsymbol{y}_t) \overset{a}{\sim} \text{FAIR}(N,H)$ entails an approximation error due to $N,H < +\infty$; asymptotics should posit $N_T, H_T \to \infty$, as $T \to \infty$.

(2) Latent variable problem: the innovations $(\boldsymbol{\varepsilon}_t)$ are typically not directly observable; direct VMA-estimation would rely on the fundamentality of $(\boldsymbol{\varepsilon}_t)$ to approximate it with in-sample forecast errors by initializing a recursion with $\{\boldsymbol{\varepsilon}_t\}_{t=-H+1}^{0} = \mathbf{0}$.

For reasons of scope, I ignore these two complications by making two very strong assumptions:

(1) $(\boldsymbol{y}_t) \sim \text{FAIR}(N,H)$, $N,H \in \mathbb{N}_0$ where the basis function family is left implicit as some mapping $\mathbb{R}^p \supset \Theta \to \mathbb{R}^{k^2(H+1)} : \boldsymbol{\theta} \mapsto (\underline{\boldsymbol{\Psi}}_h(\boldsymbol{\theta}))_{h\in\{0,\ldots,H\}}$, with $\boldsymbol{\theta}_0$ as the true, i.e. data-

[14] I.e. $\mathbb{E}\{\mathbb{E}\{X\,|\,\mathcal{F}\}\,|\,\mathcal{G}\} = \mathbb{E}\{X\,|\,\mathcal{G}\}$ for σ-fields $\mathcal{G} \subseteq \mathcal{F}$ (cf. Klenke, 2013, Theorem 8.14).

generating, parameter:

$$\boldsymbol{y}_t = \underline{\boldsymbol{\Psi}}_0(\boldsymbol{\theta}_0)\boldsymbol{\varepsilon}_t + \sum_{h=1}^{H} \underline{\boldsymbol{\Psi}}_h(\boldsymbol{\theta}_0)\boldsymbol{\varepsilon}_{t-h},\ t \in \mathbb{N}$$

(2) $\boldsymbol{\varepsilon}_t \begin{cases} = 0 \text{ a.s. } \forall t \leq 0, \\ \overset{\text{i.i.d.}}{\sim} (\boldsymbol{0}, \underline{\boldsymbol{I}})\ \forall t \in \mathbb{N} \end{cases}$ (this is a trick suggested by Wooldridge (1994)).

Now notice that by construction the innovations are known:

$$\boldsymbol{\varepsilon}_{-H+1} = \ldots = \boldsymbol{\varepsilon}_0 = \boldsymbol{0} \implies \boldsymbol{\varepsilon}_1 = \underline{\boldsymbol{\Psi}}_0(\boldsymbol{\theta}_0)^{-1}\boldsymbol{y}_1,\ \boldsymbol{\varepsilon}_t = \underline{\boldsymbol{\Psi}}_0(\boldsymbol{\theta}_0)^{-1}\left[\boldsymbol{y}_t - \sum_{h=1}^{H} \underline{\boldsymbol{\Psi}}_h(\boldsymbol{\theta}_0)\boldsymbol{\varepsilon}_{t-h}\right], \tag{VI.1}$$

but also, $(\varepsilon)_{t\in\mathbb{N}}$ is i.i.d. white noise by construction. This means that we can use (VI.1) for estimation, whereas for asymptotics we exploit that $(\boldsymbol{y}_t)_{t\geq H+1}$ is strictly stationary and ergodic. Now the $\mathcal{N}$-QMLE is defined exactly as in (V.1), with[15]

$$\boldsymbol{x}_t := (\boldsymbol{\varepsilon}_{t-1}^\top, \ldots, \boldsymbol{\varepsilon}_{t-H}^\top)^\top, \quad \boldsymbol{\mu}_t(\boldsymbol{x}_t, \boldsymbol{\theta}) \equiv \sum_{h=1}^{H} \underline{\boldsymbol{\Psi}}_h(\boldsymbol{\theta})\boldsymbol{\varepsilon}_{t-h}, \quad \underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t, \boldsymbol{\theta}) \equiv \underline{\boldsymbol{\Psi}}_0(\boldsymbol{\theta})\underline{\boldsymbol{\Psi}}_0(\boldsymbol{\theta})^\top,$$

and we can make the following observations:

(i) Assuming Θ is compact, $q := -\log f$ and L_T satisfy the Assumptions (A0) and (A1) in Theorem 2 (we can use the QMLE-notation from Section II and note that the conditional moments are time-invariant functions, thus the density terms are time-invariant).

(ii) $L_T(\boldsymbol{\theta}) = \underbrace{\frac{1}{T}\sum_{t=1}^{H} q(\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{\theta})}_{=O_p(T^{-1})} + \underbrace{\frac{T-H}{T}}_{=1+o(1)} \cdot \underbrace{\frac{1}{T-H}\sum_{t=H+1}^{T} q(\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{\theta})}_{=:\widehat{Q}_T^H},$
where the boundedness of the first term is a consequence of q a.s. cont. and Θ compact; Birkhoff's ergodic theorem now yields a pointwise LLN for $L_T(\boldsymbol{\theta})$; Then, provided an integrable bound on $|\log f|$ exists (not explored here), Lemma 2 ensures that a WULLN holds: $\max_{\boldsymbol{\theta}\in\Theta} |L_T(\boldsymbol{\theta}) - \mathbb{E}\{q(\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{\theta})\}| \xrightarrow{\text{p}} 0$.

(iii) We have by Proposition 5 an identifiably unique maximizer $\boldsymbol{\theta}_0$ of $\boldsymbol{\theta} \mapsto \mathbb{E}\{q(\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{\theta})\}$.

[15] I assume that a suitable identification criterion has been chosen, that allows to infer $\underline{\boldsymbol{\Psi}}_0$ from $\underline{\boldsymbol{\Psi}}_0\underline{\boldsymbol{\Psi}}_0^\top$. Otherwise, maximizing the likelihood is endangered by the possibility that for the maximizer $\boldsymbol{\theta}_0$ there exists a $\boldsymbol{\vartheta}$, with $(\underline{\boldsymbol{\Psi}}_h(\boldsymbol{\theta}_0))_{h\in\{0,\ldots,H\}} = (\underline{\boldsymbol{\Psi}}_h(\boldsymbol{\vartheta}) \cdot \underline{\boldsymbol{Q}})_{h\in\{0,\ldots,H\}}$ for $\underline{\boldsymbol{Q}}$ a rotation matrix; In such a case, the likelihood of both MA-parameters is the same and the optimizer is not unique. This is in principle not a problem for asymptotics, but undesirable for practical applications.

Now (i)–(iii) imply, by Theorem 3, that $\hat{\boldsymbol{\theta}} \xrightarrow{\mathrm{p}} \boldsymbol{\theta}_0$, i.e. that estimation of a FAIR-model using $\mathcal{N}$-QMLE is indeed consistent if the DGP is a FAIR-process.

For asymptotic normality we can check the Assumptions in Theorem 4 one by one:

(A0)–(A3)' verified already

(A5) (a) follows analogously to the WULLN above – it only remains to be verified whether we have an integrable bound; (b) follows from the MDS-CLT[16], and the facts that the first and second moments are correctly specified and that the model $(\boldsymbol{\mu}, \underline{\boldsymbol{\Omega}})$ is dynamically complete.

Under these conditions, the FAIR-$\mathcal{N}$-QMLE-estimator for the FAIR-process $(\boldsymbol{y}_t)$ is asymptotically normal; we can estimate the variances as outlined above, noticing that the score has zero autocorrelation, since it is an MDS.

VII. Conclusion

As for the cross-section case, QMLE-/M-estimation for time series represents a useful framework to derive asymptotic properties of specialized estimators. In this report, I reproduce the key results from this literature and show how they may be applied in the specific use-case of simplified FAIR estimation. As with direct VMA-estimation, showing consistency and asymptotic normality of the FAIR-estimator is not straightforward. While the latent-variable-problem can perhaps be overcome by relying on the fundamentality of the innovations (in principle, the same techniques as for direct VMA-estimation can be applied), the approximation problem is probably best dealt with on the population level. That is, one could show that the FAIR estimates converge in probability against a sequence of population parameters $\boldsymbol{\theta}_T^*$ (with dimension growing with T) which parametrize the KL-best approximation of a FAIR-process to the true VMA-model. The fact that we only have an approximation means that unfortunately we cannot rely on the Bollerslev and Wooldridge (1992)-approach, but have to define and analyze a new sequence of population objectives. Then, convergence of the $\boldsymbol{\theta}_T^*$-induced MA-parameters against the true MA-parameters could be addressed by the proof given by Barnichon and Matthes (2018) in their Appendix.

[16] Applied to $\sqrt{T-H}\frac{1}{T-H}\sum_{t=H+1}^{T} \boldsymbol{s}_t(\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{\theta}_0)$; This additionally requires fourth moments to exist, $\mathbb{E}\{\|\boldsymbol{s}_t(\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{\theta}_0)\|_4^4\} < \infty$, and the second moments to stabilize, i.e. $\operatorname{plim}_{T\to\infty} \frac{1}{T-H}\sum_{t=H+1}^{T} \boldsymbol{s}_t(\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{\theta}_0)\boldsymbol{s}_t(\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{\theta}_0)^\top = \lim_{T\to\infty} \operatorname{Var}\left(\sqrt{T-H}\frac{1}{T-H}\sum_{t=H+1}^{T} \boldsymbol{s}_t(\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{\theta}_0)\right)$ to both exist. (By strict stationarity and ergodicity of $(\boldsymbol{y}_t, \boldsymbol{x}_t)$, and uncorrelatedness of $(\boldsymbol{s}_t(\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{\theta}_0))$, the ergodic theorem implies that existence of fourth moments of the score is sufficient for the stabilization of second moments – the second condition is redundant.)

References

Barnichon, R., Debortoli, D., and Matthes, C. (2021). Understanding the Size of the Government Spending Multiplier: It's in the Sign. *The Review of Economic Studies*. rdab029.

Barnichon, R. and Matthes, C. (2018). Functional Approximation of Impulse Responses. *Journal of Monetary Economics*, 99:41–55.

Bollerslev, T. and Wooldridge, J. M. (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econometric Reviews*, 11(2):143–172.

Boyd, S., Boyd, S. P., and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Brockwell, P. J. and Davis, R. A. (2009). *Time series: Theory and Methods*. Springer Science & Business Media.

Davidson, J. (1994). *Stochastic Limit Theory: An Introduction for Econometricians*. OUP Oxford.

Domowitz, I. and White, H. (1982). Misspecified models with dependent observations. *Journal of Econometrics*, 20(1):35–58.

Klenke, A. (2013). *Probability theory: a Comprehensive Course*. Springer Science & Business Media.

Magnus, J. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley and Sons, New York.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, pages 1–25.

White, H. (1996). *Estimation, Inference and Specification Analysis*. Number 22. Cambridge University Press.

White, H. (2014). *Asymptotic Theory for Econometricians*. Emerald Publishing Limited.

White, H. and Domowitz, I. (1984). Nonlinear regression with dependent observations. *Econometrica*, pages 143–161.

Wooldridge, J. M. (1986). *Asymptotic Properties of Econometric Estimators*. PhD thesis, University of California, San Diego.

Wooldridge, J. M. (1994). Estimation and inference for dependent processes. *Handbook of Econometrics*, 4:2639–2738.

Appendix

A. Derivation of equation (V.2)

The t-th summand of the objective is

$$\log f_t(\boldsymbol{w}_t, \boldsymbol{\theta}) = -\frac{1}{2} \cdot \left[\log\det \underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t, \boldsymbol{\theta}) + (\boldsymbol{y}_t - \boldsymbol{\mu}_t(\boldsymbol{x}_t, \boldsymbol{\theta}))^\top \underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t, \boldsymbol{\theta})^{-1} (\boldsymbol{y}_t - \boldsymbol{\mu}_t(\boldsymbol{x}_t, \boldsymbol{\theta}))\right] + t.i.p.,$$

which itself consists of two summands. Differentiation of the first summand with respect to $\boldsymbol{\theta}$ yields

$$\frac{\partial}{\partial \boldsymbol{\theta}} \left\{\log\det \underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t, \boldsymbol{\theta})\right\} = -\frac{1}{2} \frac{\partial (\operatorname{vec}(\underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t, \boldsymbol{\theta})))^\top}{\partial \boldsymbol{\theta}} \left(\underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t, \boldsymbol{\theta})^{-1} \otimes \underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t, \boldsymbol{\theta})^{-1}\right) \cdot \operatorname{vec}\left[\underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t, \boldsymbol{\theta})\right].$$

Note that here we take the transpose of the Jacobian to be the derivative. This is obtained as follows: for some matrix $\underline{\boldsymbol{H}}$, we have $\frac{\partial \log\det \underline{\boldsymbol{H}}}{\partial \underline{\boldsymbol{H}}} := \left[\frac{\partial}{\partial [\underline{\boldsymbol{H}}]_{i,j}} \log\det \underline{\boldsymbol{H}}\right]_{i,j \in \{1,\ldots,\dim \underline{\boldsymbol{H}}\}^2} = (\underline{\boldsymbol{H}}^{-1})^\top$ (See Boyd et al., 2004, section A.4.1). Thus, $\frac{\partial \log\det \underline{\boldsymbol{\Omega}}}{\partial (\operatorname{vec} \underline{\boldsymbol{\Omega}})^\top} = (\operatorname{vec}(\underline{\boldsymbol{\Omega}}^{-1}))^\top$ and using the chain rule delivers $\frac{\partial}{\partial \boldsymbol{\theta}^\top} \{\log\det \underline{\boldsymbol{\Omega}}\} = (\operatorname{vec}(\underline{\boldsymbol{\Omega}}^{-1}))^\top \frac{\partial \operatorname{vec} \underline{\boldsymbol{\Omega}}}{\partial \boldsymbol{\theta}^\top}$, which yields the above expression after transposing and using $\operatorname{vec}(\underline{\boldsymbol{A}}\underline{\boldsymbol{B}}\underline{\boldsymbol{C}}) = (\underline{\boldsymbol{C}}^\top \otimes \underline{\boldsymbol{A}}) \operatorname{vec} \underline{\boldsymbol{B}}$ for conformable matrices.
Differentiation of the second summand with respect to $\boldsymbol{\theta}$ yields

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}} \left\{(\boldsymbol{y}_t - \boldsymbol{\mu}_t(\boldsymbol{x}_t, \boldsymbol{\theta}))^\top \underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t, \boldsymbol{\theta})^{-1} (\boldsymbol{y}_t - \boldsymbol{\mu}_t(\boldsymbol{x}_t, \boldsymbol{\theta}))\right\} &= \frac{\partial \boldsymbol{\mu}_t(\boldsymbol{x}_t, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t, \boldsymbol{\theta})^{-1} \boldsymbol{u}_t(\boldsymbol{x}_t, \boldsymbol{\theta}) \\ &\quad + \frac{1}{2} \frac{\partial (\operatorname{vec}(\underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t, \boldsymbol{\theta})))^\top}{\partial \boldsymbol{\theta}} \left(\underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t, \boldsymbol{\theta})^{-1} \otimes \underline{\boldsymbol{\Omega}}_t(\boldsymbol{x}_t, \boldsymbol{\theta})^{-1}\right) \cdot \operatorname{vec}\left[\boldsymbol{u}_t(\boldsymbol{x}_t, \boldsymbol{\theta}) \boldsymbol{u}_t(\boldsymbol{x}_t, \boldsymbol{\theta})^\top\right]. \end{aligned} \tag{A.1}$$

That we get a sum of two terms is a consequence of the multivariate product-rule. The form of the first summand is again a direct consequence of the product rule and the symmetry of the differentiated term. The second term is not so straightforward to see. Consider first some differentiable function $\underline{\boldsymbol{X}} : \mathbb{R} \to \mathbb{R}^{p \times p}$. Then the product rule delivers

$$\begin{aligned} & \frac{\partial}{\partial t} \{\underline{\boldsymbol{X}}(t) \cdot \underline{\boldsymbol{X}}(t)^{-1}\} = \frac{\partial}{\partial t} \{\underline{\boldsymbol{I}}\} \\ \iff\ & \underline{\boldsymbol{X}}(t) \frac{\partial \underline{\boldsymbol{X}}(t)^{-1}}{\partial t} + \frac{\partial \underline{\boldsymbol{X}}(t)}{\partial t} \underline{\boldsymbol{X}}(t)^{-1} = \underline{\boldsymbol{0}} \\ \implies\ & \frac{\partial \underline{\boldsymbol{X}}(t)^{-1}}{\partial t} = -\underline{\boldsymbol{X}}(t)^{-1} \frac{\partial \underline{\boldsymbol{X}}(t)}{\partial t} \underline{\boldsymbol{X}}(t)^{-1}. \end{aligned} \tag{A.2}$$

We now find $\frac{\partial \underline{\boldsymbol{X}}^{-1}}{\partial [\underline{\boldsymbol{X}}]_{i,j}}$ by thinking of $\underline{\boldsymbol{X}}$ as a map $t \mapsto [\mathbb{1}\{(\ell, h) = (i, j)\}t + \mathbb{1}\{(\ell, h) \neq (i, j)\}[\underline{\boldsymbol{X}}]_{\ell,h}]_{\ell,h\in\ldots}$ which we want to differentiate at $t = [\underline{\boldsymbol{X}}]_{i,j}$. Using (A.2) we obtain

$$
\begin{aligned}
\frac{\partial \underline{\boldsymbol{X}}^{-1}}{\partial [\underline{\boldsymbol{X}}]_{i,j}} &= -\underline{\boldsymbol{X}}^{-1}[\mathbb{1}\{(\ell, h) = (i, j)\}]_{\ell,h\in\ldots}\underline{\boldsymbol{X}}^{-1} \\
\implies \frac{\partial \operatorname{vec}(\underline{\boldsymbol{X}}^{-1})}{\partial [\underline{\boldsymbol{X}}]_{i,j}} &= \operatorname{vec}\left(-\underline{\boldsymbol{X}}^{-1}[\mathbb{1}\{(\ell, h) = (i, j)\}]_{\ell,h\in\ldots}\underline{\boldsymbol{X}}^{-1}\right) \\
&= -\left((\underline{\boldsymbol{X}}^{-1})^\top \otimes \underline{\boldsymbol{X}}^{-1}\right) \operatorname{vec}\left([\mathbb{1}\{(\ell, h) = (i, j)\}]_{\ell,h\in\ldots}\right) \\
&= \left[-(\underline{\boldsymbol{X}}^{-1})^\top \otimes \underline{\boldsymbol{X}}^{-1}\right]_{:,p\cdot(i-1)+j}
\end{aligned}
$$

and thus $\frac{\partial \operatorname{vec} \underline{\boldsymbol{X}}^{-1}}{\partial (\operatorname{vec} \underline{\boldsymbol{X}})^\top} = -\left((\underline{\boldsymbol{X}}^{-1})^\top \otimes \underline{\boldsymbol{X}}^{-1}\right)$. This can be applied as follows:

$$
\begin{aligned}
\underbrace{\frac{\partial}{\partial (\operatorname{vec} \underline{\boldsymbol{\Omega}})^\top} \underbrace{\boldsymbol{u}^\top \underline{\boldsymbol{\Omega}}^{-1} \boldsymbol{u}}}_{=\operatorname{vec}(\boldsymbol{u}^\top \underline{\boldsymbol{\Omega}}^{-1}\boldsymbol{u})=(\boldsymbol{u}^\top \otimes \boldsymbol{u}^\top)\operatorname{vec}(\underline{\boldsymbol{\Omega}}^{-1})=\operatorname{vec}(\boldsymbol{u}\boldsymbol{u}^\top)^\top \operatorname{vec}(\underline{\boldsymbol{\Omega}}^{-1})} &= -\operatorname{vec}(\boldsymbol{u}\boldsymbol{u}^\top)^\top(\underline{\boldsymbol{\Omega}}^{-1} \otimes \underline{\boldsymbol{\Omega}}^{-1}) \\
&= -\left[(\underline{\boldsymbol{\Omega}}^{-1} \otimes \underline{\boldsymbol{\Omega}}^{-1})\operatorname{vec}(\boldsymbol{u}\boldsymbol{u}^\top)\right]^\top .
\end{aligned}
$$

Using the chain rule one last time delivers the second summand of the derivative in (A.1). ■